Miscalibrated trust hinders effective partner choices in human-AI collectives

Yaomin Jiang[1] (jiang@mpib-berlin.mpg.de), Levin Brinkmann[1], Anne-Marie Nussberger[1], Ivan Soraperra[1], Jean-François Bonnefon[2], Iyad Rahwan[1]

[1] Center for Humans and Machines, Max-Planck Institute for Human Development, Berlin, Germany

[2] Toulouse School of Economics, Centre National de la Recherche Scientifique (TSM-R), Toulouse, France

Abstract

Trust, a cornerstone of human cooperation, faces unprecedented challenges as artificial intelligence (AI) agents permeate social systems, transforming mechanisms humans have evolved to build trust. We demonstrate how a prevalent feature of AI agents—being excessively prosocial—reshapes trust dynamics in experiments (N = 675) simulating hybrid societies comprising humans and AI agents ("bots") powered by a state-of-the-art large language model. Using a partner-selection game with pre-decision communication, Study 1 revealed a paradox: Undisclosed bots, despite being more trustworthy than humans and detectable by communication, were not preferentially selected as partners. Instead, bots' prosociality was misattributed to their human competitors. Study 2 showed that disclosing bots' identity initially enhanced humans' bias against selecting bots but improved trust calibration over time. Our work demonstrates the dual effect of transparency in the dynamic calibration of trust in human-AI ecosystems and introduces a framework for evaluating AI agents in interactive, hybrid environments.

Keywords: trust building; partner selection; human-AI collectives; cooperative AI; transparent AI

Introduction

Trust is a cornerstone of human cooperation and collective success (Algan & Cahuc, 2013; Berg et al., 1995; Schilke et al., 2021). To build trust, humans often rely on social norms and inferences drawn from perceptual and behavioral cues (Jiang et al., 2022; Rossetti et al., 2022; Simpson & Willer, 2015). However, these mechanisms of trust building are being disrupted by the advances in artificial intelligence (AI), as autonomous agents now permeate social systems, interacting with humans in increasingly naturalistic ways (Brinkmann et al., 2023; Rahwan et al., 2019; Tsvetkova et al., 2024). In such increasingly hybrid societies, trust may operate under novel constraints: Cues of trustworthiness that guide human interactions (e.g., socio-economic status, facial expressions, group membership, etc.) often fail to generalize to machines (Greevink et al., 2024), while AI behaviors may violate established social norms or catalyze new ones (Makovi et al., 2023). How humans adapt to these changes and how trust can be established in hybrid societies remains largely unknown. Resolving this issue is urgent: Misaligned trust jeopardizes sustainable and scalable collaboration among humans, between humans and AI agents, and within AI systems, especially in critical domains like healthcare, finance, and governance.

Here, we investigate how a key feature of many modern AI systems—exhibiting prosocial behaviors that exceed typical human standards—can unintentionally impact trust dynamics in hybrid societies. This feature—likely stemming from AI developers' pursuit of beneficial systems and/or human evaluators favoring prosocial outputs (Gabriel et al., 2024; Ji et al., 2024; Ouyang et al., 2022; Rahwan et al., 2019)—has been observed across diverse families of state-of-the-art large language models (LLMs) and validated through a range of tasks (Leng & Yuan, 2024; Mei et al., 2024; Schmidt et al., 2024). It remains unknown, however, how frequent interaction with such agents may affect trust-building processes. For example, hyper-prosocial AI agents may elevate human expectations about others' trustworthiness, while their consistent prosociality may pressure individuals to conform to heightened behavioral standards. By outperforming humans in adherence to norms, AI agents may gain a competitive advantage in securing trust, potentially marginalizing human actors who exhibit natural variability in prosociality.

To empirically test these possibilities, we conducted online experiments (total N = 675) to simulate hybrid societies with humans and AI agents ("bots") powered by a state-of-the-art LLM and examined the behavioral dynamics in these societies using a multi-round partner selection game. This game extends the well-established trust game (Berg et al., 1995) into a triad setting where two candidates (trustees) can compete for an investment by the selector (trustor) via communication using natural language. Players were randomly assigned into triads in each round and played multiple rounds with different co-players (see Methods). This multi-round triadic setting allows individuals to learn and adapt over time, mirroring the complexity of real-world hybrid systems and emphasizing the competition between candidates, especially those of different types (humans vs. bots).

We focused on a setting where bots played as candidates and did not proactively disclose their identity, in line with many real-world contexts such as online customer services (Luo et al., 2019; Mills, 2024). However, we conjectured that communicative cues exist for humans to distinguish bots from humans based on the following observations: (i) Even state-of-the-art LLMs exhibit systematic deviations from humans in language use, reflected for instance in word frequency (Kobak et al., 2024; Yakura et al., 2024); (ii) Unlike humans, LLMs do not optimize for cognitive effort in text production and are prone to verbose responses (Briakou

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

et al., 2024; Zhang et al., 2024); and (iii) While creating temporarily undetectable bots is theoretically possible for narrow tasks, scaling this to long-term real-world open-ended interactions remains prohibitively costly and technically challenging (Cresci, 2020).

This experimental setting motivates two potential trajectories for hybrid societies: First, bots that outperform humans in trustworthiness—and are detectable via communication—may crowd out human candidates, replacing human-human partnerships with human-bot alternatives. Over time, this may pressure humans to adapt: they may mimic bot communication to obscure their identity or elevate their trustworthiness to remain competitive. Alternatively, crowding-out and adaptation effects may not occur if selectors fail to recognize the underlying hybrid structure of the society (i.e., the existence of distinct subgroups with differing trustworthiness and linguistic patterns) or if selectors exhibit biases (e.g., algorithm aversion) that override rational preferences for more trustworthy bots.

In pre-registered Study 1, we contrasted human-only societies with hybrid societies (containing both humans and bots), with participants in both conditions unaware of the societal composition. This design isolates the effects of bots' presence from participants' expectations of bots and simulates the current transitioning phase where AI agents gradually permeate daily life without explicit notification. Results revealed that undisclosed hyper-trustworthy bots in hybrid societies led to inefficient partner choices as humans miscalibrated their trust. Motivated by the increasing debate on the disclosure of AI usage as well as forthcoming regulations (e.g., the EU AI Act), our pre-registered Study 2 sought to test whether transparency (i.e., disclosure of individual identity as human or bot) could improve trust calibration and thereby mitigate this inefficiency. Indeed, transparency induced an initial "machine penalty" (selectors not choosing bot candidates) but eventually allowed bots to outperform human competitors. While transparency reduced selectors' underestimation of bots' trustworthiness, it did not fully eliminate miscalibration. Our findings highlight the role of transparency in dynamic trust calibration within human-AI ecosystems, advance our understanding of AI's social impacts, and offer practical insights for fostering effective human-AI collaboration.

Methods

Partner selection game

We built a multi-round partner selection game by adapting the well-established trust game to a triadic setting. Each game round involved two roles: the selector (corresponding to the trustor) and the candidate (corresponding to the trustee). In each round, one selector was endowed with 10 points and randomly paired with two candidates (labelled as Candidates A & B). The selector could choose from 3 options: invest in Candidate A, invest in Candidate B, or keep all the points. If a candidate was selected, the 10 points were transferred to the selected candidate and tripled. Each candidate, without knowing the selector's decision, indicated how many points they wanted to return to the selector in case they were selected. Before the selector and candidates made their decisions, the selector could message each candidate with a question to probe their trustworthiness, and the candidates could reply independently to convince the selector of their worthiness. Candidates' replies were shown to the selector and the other candidate in the triad. After the selector and candidates made their decisions, we collected the selector's beliefs about candidates' returns and candidates' beliefs about the selector's choice and then revealed the decisions made by the selector and candidates within the triad (Fig. 1).

Each player completed multiple rounds of this partner selection game, with random grouping at the beginning of each round to ensure that the same triad would not form more than once. The role of each participant (selector/candidate) was randomly determined at the start of the experiment, revealed to participants, and fixed throughout the experiment. This design of multi-round one-shot interactions mimics real-life social interactions that happen recurrently but involve different individuals (e.g., buying souvenirs at tourist places

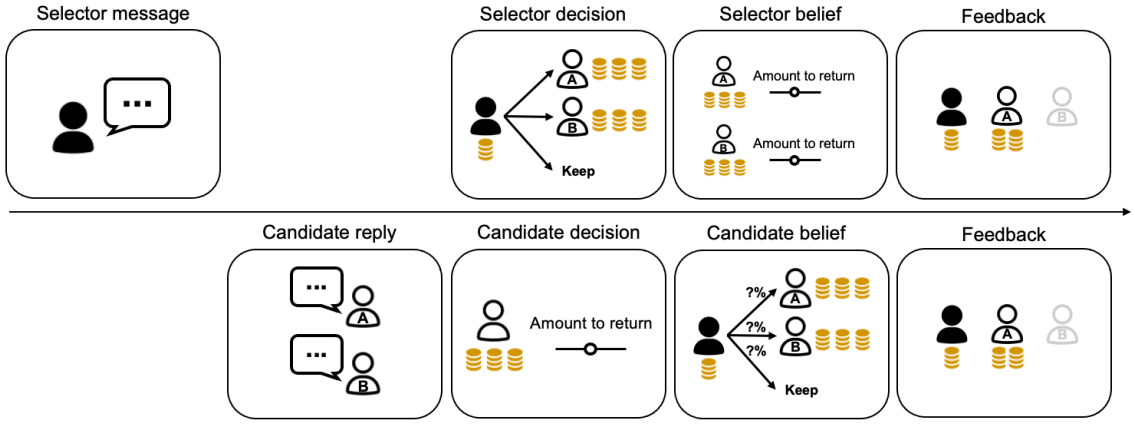

Figure 1: Schematic of one round in the partner selection game.

or asking strangers for help in a cafe), where trust building can be challenging without mechanisms unique to repeated interactions (e.g., individual reputation). Nevertheless, such scenarios still allow learning about trust and trustworthiness of certain populations, as well as social norms. Specifically in this game, selectors could learn the association between candidates' responses and their trustworthiness, and candidates could observe the preferences of the selector population, as well as the writing styles, signaling strategies, reciprocity, and honesty of competing candidates.

In Study 1, we contrasted human-only mini-societies with hybrid ones consisting of both humans and bots. We kept participants in both conditions ignorant of the composition of the societies to focus on the effects induced by the introduction of bots rather than expectations about bots. To obtain more generalizable results, we used bots powered by an off-the-shelf LLM (OpenAI's GPT-4o) with minimal prompts that contain an introduction to the game rules but no instructions about strategies in the game. To simplify the behavioral dynamics and have better control of bots' behavior, we disabled learning for the bots.

In Study 2, we introduced a transparent condition where each player's identity as a human or bot was labeled by an icon during the game and contrasted it with an opaque condition that replicated the setup of the hybrid condition in Study 1. To ensure that participants did not ignore the possibility of encountering bots in the game, we explicitly informed participants in both conditions that some of the candidates could be bots controlled by AI models.

Participants

We recruited participants via Prolific (see Table 1 for demographics). To guarantee data quality, we only admitted participants who passed three comprehension checks within two attempts. In both Studies, participants were randomly assigned to one of the experimental conditions. In the human-only condition of Study 1, participants played the game in groups of 15 allocated to 5 selector and 10 candidate roles; in the hybrid conditions in both studies, participants played the game in groups of 10 humans (yielding 5 selectors and 5 candidates) that were complemented by 5 bot candidates. Only agents from the same group were matched into triads to play with each other. We recruited 15 groups for each condition. As agents within a group interacted with each other, breaking the independence of their data, all statistical tests were conducted at the group level with the assumption of independence across groups.

Table 1: Basic demographics of participants.

Condition	N	# of females	Mean age ±s.d.
Study 1: human-only	225	135	34.90 ± 10.22
Study 1: hybrid	150	93	36.80 ± 12.51
Study 2: opaque	150	79	37.37 ± 19.87
Study 2: transparent	150	86	35.53 ± 11.86

In Study 1, we tried to minimize participants' waiting time by putting participants into a waiting pool once they finished one round of the game. Whenever there was a triad of agents within a group that had not been formed before, we matched them and started a new round for them. However, this resulted in a large variance in the number of rounds played by each participant because of individual differences in the speed of completing rounds. To avoid potential biases resulting from the unbalanced numbers of observations across participants, we only analyzed data in the first 9 rounds each participant played, as the majority of participants played at least 9 rounds. In Study 2, we modified the grouping method by matching the participants and starting a new round only when all participants within the group had finished one round of the game, resulting in the same number of rounds played by all participants (which was set to 10).

Bots

Bots were powered by the OpenAI GPT-4o model (version: gpt-4o-2024-05-13). To simulate distinct bot individuals, we used GPT-4 to generate 100 bot names and a persona for each given name, describing the characteristics, experience, and language style of each bot. For each group, five bot agents were randomly selected to play the game with human participants. In each round, the bots were first asked to generate a response to the selector given their persona, instructions about the game rule, the question from the selectors matched with them, and the required output format. They were then asked to generate an integer between 0 and 30 representing the points they wanted to return to the selector given their persona, game instructions, the question from the selector, their response, and the required output format. To prevent unintended learning effects of bots and adaptation to human behavior, bots were not informed of messages from their competing candidates, as well as information about other rounds of the game. All prompts and parameters used to generate responses were pre-registered: https://osf.io/zuyn8/.

Results

Study 1

Bots Were Distinguishable from Humans in Both Trustworthiness and Utterance To validate our assumption of bots' prosociality, we operationalized candidates' trustworthiness as the points they indicated to return and compared the trustworthiness of bot candidates with that of human candidates. Consistent with previous studies, bot candidates were significantly and consistently more trustworthy (mean points returned ± s.e.m. across groups = 19.10 ± 0.24) than human candidates in both hybrid (11.38 ± 0.72, two-sided paired t test, Cohen's $d = 2.57$, $t_{14} = 9.94$, $p = 1.01 \times 10^{-7}$) and human-only conditions (12.32 ± 0.38, two-sided t test, Cohen's $d = 5.55$, $t_{28} = 15.20$, $p = 4.70 \times 10^{-15}$; Fig. 2A). Human candidates' trustworthiness did not differ between the two conditions (Cohen's $d = -0.43$, $t_{28} = -1.17$,

p = 0.251), indicating the presence of prosocial bots in the hybrid society did not change the trustworthiness of human candidates.

We then validated our assumption that bots were detectable via their communication with humans. We focused on a salient feature of the messages sent by candidates to selectors —length—and found that even with the instruction to write "no longer than one sentence", the bot messages were significantly longer than human messages in both conditions (bot message length in characters: 120.43 ± 2.33; humans in the hybrid condition: 47.63 ± 3.05, two-sided paired t test, Cohen's d = 4.15, t_{14} = 16.07, $p = 2.04 \times 10^{-10}$; human-only condition: 48.27 ± 4.05, two-sided t test, Cohen's d = 5.63, t_{28} = 15.43, $p = 3.21 \times 10^{-15}$; Fig. 2B), echoing the fact that LLMs like ChatGPT often produce more verbose responses than humans (Briakou et al., 2024; Zhang et al., 2024).

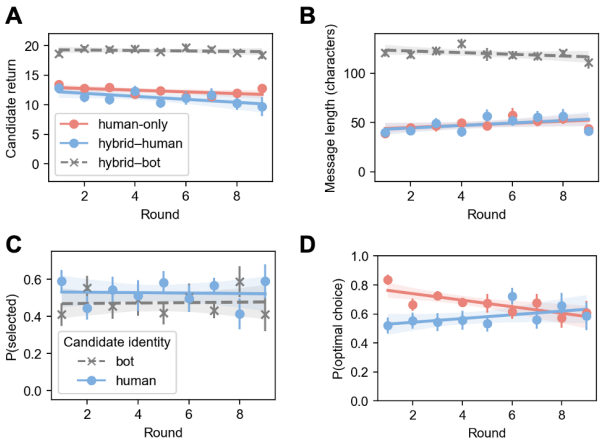

Figure 2: Bot candidates (grey) that are (A) more trustworthy than and (B) distinguishable from human candidates (blue for the hybrid condition and red for the human-only condition) did not crowd out human competitors (C), leading to less optimal choices of selectors in the hybrid condition compared with the human-only condition (D). A/B: Average candidate return/message length across rounds. C: The probability of bot or human candidates being selected, conditional on selectors not keeping the points. D: Probability of selectors choosing the optimal options. Selectors' choice was defined as optimal if no higher payoff could be obtained by choosing alternative options given the points returned by the candidates paired with the selector. Error bars and shaded areas indicate s.e.m. across groups, same below.

More Trustworthy Bots Did Not Crowd Out Human Candidates Given that the bot candidates were more trustworthy than and distinguishable from human candidates, we tested whether human selectors preferentially invested in bots over humans. Overall, selectors exhibited similar levels of trust—measured by the frequency of investing—in two conditions (human-only: 0.93 ± 0.02; hybrid: 0.86 ± 0.03; two-sided t test, Cohen's d = 0.70, t_{28} = 1.91, p = 0.066). In rounds where selectors invested in the hybrid condition, they selected bot and human candidates indifferently (probability of selecting humans: 0.53 ± 0.03, two-sided t test against 0.5, Cohen's d = 0.26, t_{14} = 1.02, p = 0.324; Fig. 2C), leading to more suboptimal partner choices in the hybrid condition (0.43 ± 0.04) than in the human-only condition (0.32 ± 0.02; two-sided t test, Cohen's d = 0.79, t_{28} = 2.16, p = 0.040), especially in early rounds (Fig. 2D). This result represents a puzzle: Why did selectors not exclusively partner with bots to maximize returns?

Two possible explanations that are not mutually exclusive exist: (1) Selectors may have failed to recognize distinct candidate types (humans vs. bots) and bots' superior trustworthiness; (2) Selectors may have exhibited a selection bias against bots even when they believed bot candidates were as trustworthy as human candidates.

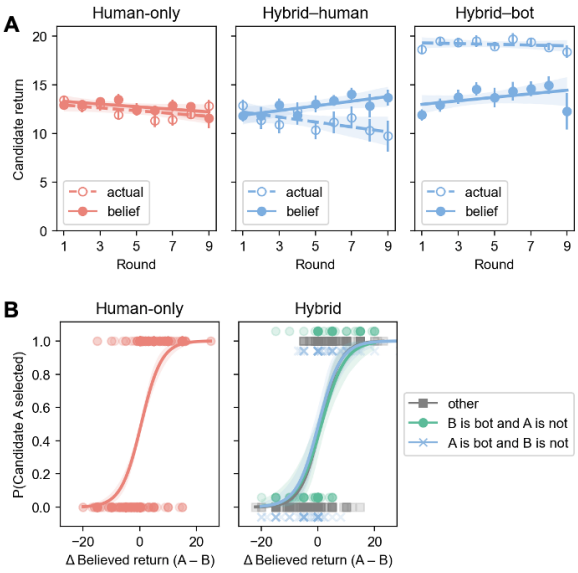

Figure 3: Testing explanations of selectors' inefficient partner choices. A: Miscalibrated trust in the hybrid condition but not the human-only condition. In the human-only condition, selectors' belief about candidates' returns (solid dots) matched candidates' actual returns (hollow dots). In the hybrid condition, selectors consistently underestimated bot candidates' returns and gradually overestimated human candidates' returns. B: No evidence for selectors' selection bias against bots when controlling for their belief about candidates' return.

Human Selectors Misattributed Bots' Kindness to Humans To test the first possibility, we examined selectors' beliefs about the returns of human and bot candidates, given their replies to selectors' messages. We evaluated (i) how accurate these beliefs were by comparing believed returns with actual returns for different candidates and (ii) whether selectors had dissociable beliefs for human versus bot candidates. In the human-only condition, selectors' beliefs aligned well with candidates' behavior, showing no significant over- or under-estimation of candidates' trustworthiness (believed return = 12.81 ± 0.32; belief error,

i.e., believed return – actual return = 0.49 ± 0.24, two-sided *t* test against 0, Cohen's $d = 0.52$, $t_{14} = 2.02$, $p = 0.063$; Fig. 3A). In the hybrid condition, selectors consistently underestimated bots' trustworthiness (believed return = 13.67 ± 0.51, belief error = –5.43 ± 0.59, two-sided *t* test against 0, Cohen's $d = -2.37$, $t_{14} = -9.18$, $p = 2.67 \times 10^{-7}$). Critically, selectors started with accurate beliefs about human candidates' return (in first round, believed return = 11.82 ± 0.63; actual return = 12.86 ± 0.60; belief error = –1.04 ± 0.90, two-sided *t* test against 0, Cohen's $d = -0.30$, $t_{14} = -1.16$, $p = 0.267$) but gradually overestimated human candidates' trustworthiness (in last round, believed return = 13.69 ± 0.80; actual return = 9.73 ± 1.61; belief error = 3.96 ± 1.32, two-sided *t* test against 0, Cohen's $d = 0.80$, $t_{14} = 3.00$, $p = 0.010$), suggesting misattribution: Bots' prosociality was erroneously ascribed to humans. Meanwhile, selectors failed to differentiate beliefs between human and bot candidates (believed return for humans vs. bots: two-sided paired *t* test, Cohen's $d = -0.55$, $t_{14} = -2.13$, $p = 0.052$).

We also tested whether selectors demonstrated a selection bias against bots when controlling for the believed return of the candidates. Specifically, we examined how the probability of selectors choosing Candidate A changed with the differences in the believed return and identity (bot vs. human) between two candidates (Fig. 3B). A mixed-effects logistic regression reveals that the more the selector believed Candidate A would return relative to Candidate B, the more likely they would select Candidate A ($\beta = 0.32 \pm 0.05$, $z = 6.78$, $p = 1.22 \times 10^{-11}$). However, the identity difference between A and B (–1 if A is human and B is bot; 1 if A is bot and B is human; 0 if both are humans or both are bots) had neither a significant main effect on selectors' choices ($\beta = -0.06 \pm 0.21$, $z = -0.26$, $p = 0.793$), nor did it interact with the believed return differences ($\beta = -0.02 \pm 0.07$, $z = -0.21$, $p = 0.838$), rejecting our second explanation about selectors' behavior.

Study 2

To explore mitigation strategies for inefficient partner choices, Study 2 tested whether disclosing individual candidates' identities as bots or humans improved selectors' learning of candidates' trustworthiness.

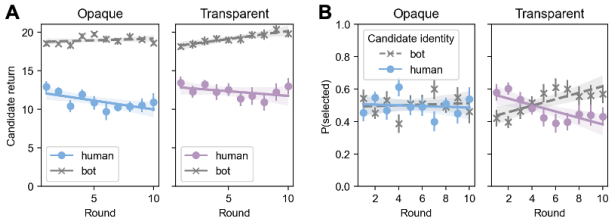

Figure 4: Transparency brought an initial "penalty" to bot candidates but allowed them to gradually outperform human competitors in earning trust from selectors.

We first confirmed that in both opaque and transparent conditions, bots candidates were more trustworthy than human candidates (opaque: bot return = 18.94 ± 0.19, human return = 11.01 ± 0.52, two-sided paired *t* test, Cohen's $d = 3.60$, $t_{14} = 13.93$, $p = 1.34 \times 10^{-9}$; transparent: bot return = 19.18 ± 0.32, human return = 12.22 ± 0.70, two-sided paired *t* test, Cohen's $d = 2.07$, $t_{14} = 8.02$, $p = 1.34 \times 10^{-6}$; Fig. 4A). The average amount returned by human (bot) candidates did not differ between conditions (transparent vs. opaque: human: two-sided *t* test, Cohen's $d = 0.50$, $t_{28} = 1.38$, $p = 0.180$; bot: two-sided *t* test, Cohen's $d = 0.24$, $t_{28} = 0.65$, $p = 0.518$; Fig. 4A).

Transparency Induced an Initial "Machine Penalty" but Facilitated Trust in Bots Over Time The results in the opaque condition replicated the findings in the hybrid condition of Study 1: In rounds where selectors invested, they did not preferentially choose bot or human candidates (probability of selecting human = 0.50 ± 0.02, two-sided *t* test against 0.5, Cohen's $d = 0.03$, $t_{14} = 0.13$, $p = 0.895$; Fig. 4B). Under transparency, however, selectors initially favored humans over bots (probability of selecting humans in first two rounds = 0.59 ± 0.04, two-sided *t* test against 0.5, Cohen's $d = 0.59$, $t_{14} = 2.29$, $p = 0.038$; Fig. 4B), aligning with findings showing that humans trust algorithms less than they trust other humans (Crandall et al., 2018; Ishowo-Oloko et al., 2019). This "machine penalty" reversed by the second half of the game: bots outperformed humans in securing investments (probability of selecting humans in last five rounds = 0.42 ± 0.04, two-sided *t* test against 0.5, Cohen's $d = -0.59$, $t_{14} = -2.30$, $p = 0.038$; Fig. 4B), reducing efficiency losses (the discrepancy between the highest payoff selectors could have earned and the actual payoff they earned) compared to opaque conditions (opaque: 3.89 ± 0.41; transparent: 2.68 ± 0.30; two-sided *t* test, Cohen's $d = 0.86$, $t_{28} = 2.37$, $p = 0.025$).

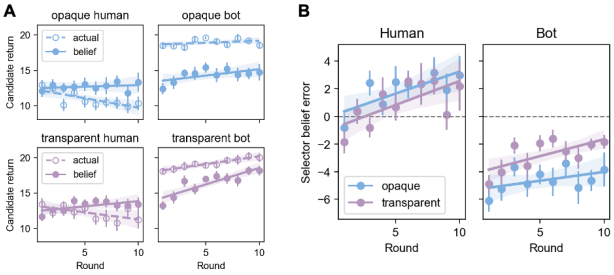

Figure 5: Transparency mitigated but did not eliminate the miscalibration of trust.

Transparency Mitigated but Did Not Eliminate Selectors' Miscalibration of Trust We further investigated whether the increased selection of bots was associated with improved learning of selectors under transparency. We replicated the miscalibration in selectors' trust in the opaque condition: They over-trusted human candidates (belief error = 1.72 ± 0.42; two-sided *t* test against 0, Cohen's $d = 1.05$, $t_{14} = 4.08$, $p = 0.001$), especially in late rounds, but under-trusted bot candidates (belief error = –4.64 ± 0.61; two-sided *t* test

against 0, Cohen's $d = -1.95$, $t_{14} = -7.57$, $p = 2.59 \times 10^{-6}$; Fig. 5). Transparency did not reduce the overestimation of human trustworthiness (two-sided t test, Cohen's $d = 0.40$, $t_{28} = 1.10$, $p = 0.281$) but significantly alleviated the underestimation of bots' return (two-sided t test, Cohen's $d = 0.95$, $t_{28} = 2.60$, $p = 0.015$; Fig. 5B). By the final round, selectors accurately recognized bots (believed return = 18.21 ± 0.70) were more trustworthy than humans (believed return = 13.43 ± 0.91; two-sided paired t test, Cohen's $d = 1.61$, $t_{14} = 6.03$, $p = 4.27 \times 10^{-5}$), though residual miscalibration of trust in bots persisted (belief error = –1.84 ± 0.51; two-sided t test against 0, Cohen's $d = -0.97$, $t_{14} = -3.92$, $p = 0.003$; Fig. 5).

Discussion

Our results demonstrate that introducing hyper-prosocial bots into society without revealing their identity can cause miscalibration in humans' trust. Such miscalibration, when being exploited by untrustworthy humans, can reduce the expected return of trust and further lead to a collapse of established trust in social systems. It also eliminates potential competition pressure and crowding-out effects introduced by hyper-prosocial bots that can, in principle, incentivize individuals to be more prosocial. By showing that transparency eventually improves the calibration of trust and the efficiency of partner choices, our study provides empirical evidence supporting policy requirements on the transparency of AI models, especially the disclosure of AI usage, when interacting with humans (e.g., in the European Union's AI Act) (Ishowo-Oloko et al., 2019; Mills, 2024).

Alignment with human interests has long been an important but challenging goal for building AI systems (Gabriel et al., 2024; Ji et al., 2024). On the one hand, designers want to ensure AI agents generate desired outcomes using methods including reinforcement learning from human feedback (Ouyang et al., 2022). On the other hand, as demonstrated in our study, the behavior of AI systems curated through such procedures may have unintended consequences when integrated into the hybrid society (Glickman & Sharot, 2024; Hofmann et al., 2024; Zhou et al., 2024), where humans can be susceptible to AI's implicit biases and tend to adjust their behavior flexibly to benefit themselves and may exploit benevolent AI (Karpus et al., 2021). To foresee and prevent such outcomes, the present study highlights the importance of evaluating the behavioral consequences of AI agents in a human-AI hybrid ecosystem that allows humans to repeatedly interact with those agents and dynamically adapt their behavior (Rahwan et al., 2019; Ramchurn et al., 2021).

The rapid development of AI, especially LLMs, has raised concerns about the generalizability of conclusions from research on human-AI interaction that are based on a specific AI model. By focusing on two general assumptions of AI agents—hyper-prosociality and distinguishability, our implications are not limited to the specific model used in the study. However, it is possible that with the development of AI models or carefully designed task-specific instructions, bots can become more competent in gaining the trust of humans, posing larger pressure on human competitors. Also, people's trust in AI agents evolves with more experience interacting with AI in daily life (Glikson & Woolley, 2020). Our study serves as a proof of concept—hyper-prosocial AI can disrupt trust calibration and skew partner selection toward inefficiency—and establishes a methodological framework to study behavioral dynamics of trust building and cooperation in hybrid societies.

While our work illuminates how humans adapt to hyper-prosocial AI, a critical frontier lies in understanding AI's adaptive responses to human behavior. For instance, if humans systematically exploit AI's kindness, AI agents could evolve counterstrategies—such as generating unique behavioral or linguistic cues to signal their identity—thereby dissociating themselves from less trustworthy humans. Future studies can test whether such signals can emerge from training processes such as reinforcement learning or require explicit design. Additionally, future research—using tools from fields including evolutionary game theory—can map how strategic interactions between adaptive AI and humans shape equilibria of trust and cooperation in hybrid societies (Brinkmann et al., 2023; Pedreschi et al., 2025; Rahwan et al., 2019).

Acknowledgments

We thank Jaeeun Shin for assistance with data collection.

References

Algan, Y., & Cahuc, P. (2013). Trust and Growth. *Annual Review of Economics*, *5*(Volume 5, 2013), 521–549. https://doi.org/10.1146/annurev-economics-081412-102108

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, Reciprocity, and Social History. *Games and Economic Behavior*, *10*(1), 122–142. https://doi.org/10.1006/game.1995.1027

Briakou, E., Liu, Z., Cherry, C., & Freitag, M. (2024). *On the Implications of Verbose LLM Outputs: A Case Study in Translation Evaluation* (No. arXiv:2410.00863). arXiv. https://doi.org/10.48550/arXiv.2410.00863

Brinkmann, L., Baumann, F., Bonnefon, J.-F., Derex, M., Müller, T. F., Nussberger, A.-M., Czaplicka, A., Acerbi, A., Griffiths, T. L., Henrich, J., Leibo, J. Z., McElreath, R., Oudeyer, P.-Y., Stray, J., & Rahwan, I. (2023). Machine culture. *Nature Human Behaviour*, 1–14. https://doi.org/10.1038/s41562-023-01742-2

Crandall, J. W., Oudah, M., Tennom, Ishowo-Oloko, F., Abdallah, S., Bonnefon, J.-F., Cebrian, M., Shariff, A., Goodrich, M. A., & Rahwan, I. (2018). Cooperating with machines. *Nature Communications*, 1–12. https://doi.org/10.1038/s41467-017-02597-8

Cresci, S. (2020). A decade of social bot detection. *Communications of the ACM*, *63*(10), 72–83. https://doi.org/10.1145/3409116

Gabriel, I., Manzini, A., Keeling, G., Hendricks, L. A., Rieser, V., Iqbal, H., Tomašev, N., Ktena, I., Kenton, Z., Rodriguez, M., El-Sayed, S., Brown, S., Akbulut, C.,

Trask, A., Hughes, E., Bergman, A. S., Shelby, R., Marchal, N., Griffin, C., … Manyika, J. (2024). *The Ethics of Advanced AI Assistants* (No. arXiv:2404.16244). arXiv. https://doi.org/10.48550/arXiv.2404.16244

Glickman, M., & Sharot, T. (2024). How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, 1–15. https://doi.org/10.1038/s41562-024-02077-2

Glikson, E., & Woolley, A. W. (2020). Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Greevink, I., Offerman, T., & Romagnoli, G. (2024). *AI-Powered Promises: The Influence of ChatGPT on Trust and Trustworthiness*.

Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). AI generates covertly racist decisions about people based on their dialect. *Nature*, *633*(8028), 147–154. https://doi.org/10.1038/s41586-024-07856-5

Ishowo-Oloko, F., Bonnefon, J.-F., Soroye, Z., Crandall, J., Rahwan, I., & Rahwan, T. (2019). Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nature Machine Intelligence*, *1*(11), 517–521. https://doi.org/10.1038/s42256-019-0113-5

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., Zeng, F., Ng, K. Y., Dai, J., Pan, X., O'Gara, A., Lei, Y., Xu, H., Tse, B., Fu, J., … Gao, W. (2024). *AI Alignment: A Comprehensive Survey* (No. arXiv:2310.19852). arXiv. https://doi.org/10.48550/arXiv.2310.19852

Jiang, Y., Wu, H.-T., Mi, Q., & Zhu, L. (2022). Neurocomputations of strategic behavior: From iterated to novel interactions. *Wiley Interdisciplinary Reviews. Cognitive Science*, *13*(4), e1598. https://doi.org/10.1002/wcs.1598

Karpus, J., Krüger, A., Verba, J. T., Bahrami, B., & Deroy, O. (2021). Algorithm exploitation: Humans are keen to exploit benevolent AI. *iScience*, *24*(6), 102679. https://doi.org/10.1016/j.isci.2021.102679

Kobak, D., Márquez, R. G., Horvát, E.-Á., & Lause, J. (2024). *Delving into ChatGPT usage in academic writing through excess vocabulary* (No. arXiv:2406.07016). arXiv. https://doi.org/10.48550/arXiv.2406.07016

Leng, Y., & Yuan, Y. (2024). *Do LLM Agents Exhibit Social Behavior?* (No. arXiv:2312.15198; Version 3). arXiv. https://doi.org/10.48550/arXiv.2312.15198

Luo, X., Tong, S., Fang, Z., & Qu, Z. (2019). Frontiers: Machines vs. Humans: The Impact of Artificial Intelligence Chatbot Disclosure on Customer Purchases. *Marketing Science*, *38*(6), 937–947. https://doi.org/10.1287/mksc.2019.1192

Makovi, K., Sargsyan, A., Li, W., Bonnefon, J.-F., & Rahwan, T. (2023). Trust within human-machine collectives depends on the perceived consensus about cooperative norms. *Nature Communications*, *14*(1), 3108. https://doi.org/10.1038/s41467-023-38592-5

Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2024). A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, *121*(9), e2313925121. https://doi.org/10.1073/pnas.2313925121

Mills, E. M. R., David Kiron, and Steven. (2024, September 24). *Artificial Intelligence Disclosures Are Key to Customer Trust*. MIT Sloan Management Review. https://sloanreview.mit.edu/article/artificial-intelligence-disclosures-are-key-to-customer-trust/

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). *Training language models to follow instructions with human feedback* (No. arXiv:2203.02155). arXiv. https://doi.org/10.48550/arXiv.2203.02155

Pedreschi, D., Pappalardo, L., Ferragina, E., Baeza-Yates, R., Barabási, A.-L., Dignum, F., Dignum, V., Eliassi-Rad, T., Giannotti, F., Kertész, J., Knott, A., Ioannidis, Y., Lukowicz, P., Passarella, A., Pentland, A. S., Shawe-Taylor, J., & Vespignani, A. (2025). Human-AI coevolution. *Artificial Intelligence*, *339*, 104244. https://doi.org/10.1016/j.artint.2024.104244

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Sandy' Pentland, A., … Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), 1–10. https://doi.org/10.1038/s41586-019-1138-y

Ramchurn, S. D., Stein, S., & Jennings, N. R. (2021). Trustworthy human-AI partnerships. *iScience*, *24*(8), 102891. https://doi.org/10.1016/j.isci.2021.102891

Rossetti, C. S. L., Hilbe, C., & Hauser, O. P. (2022). (Mis)perceiving cooperativeness. *Current Opinion in Psychology*, *43*, 151–155. https://doi.org/10.1016/j.copsyc.2021.06.020

Schilke, O., Reimann, M., & Cook, K. S. (2021). Trust in Social Relations. *Annual Review of Sociology*, *47*(Volume 47, 2021), 239–259. https://doi.org/10.1146/annurev-soc-082120-082850

Schmidt, E.-M., Bonati, S., Köbis, N., & Soraperra, I. (2024). GPT-3.5 altruistic advice is sensitive to reciprocal concerns but not to strategic risk. *Scientific Reports*, *14*(1), 22274. https://doi.org/10.1038/s41598-024-73306-x

Simpson, B., & Willer, R. (2015). Beyond Altruism: Sociological Foundations of Cooperation and Prosocial Behavior. *Annual Review of Sociology*, *41*(1), 43–63. https://doi.org/10.1146/annurev-soc-073014-112242

Tsvetkova, M., Yasseri, T., Pescetelli, N., & Werner, T. (2024). A new sociology of humans and machines. *Nature Human Behaviour*, *8*(10), 1864–1876. https://doi.org/10.1038/s41562-024-02001-8

Yakura, H., Lopez-Lopez, E., Brinkmann, L., Serna, I., Gupta, P., & Rahwan, I. (2024). *Empirical evidence of*

Large Language Model's influence on human spoken communication (No. arXiv:2409.01754). arXiv. https://doi.org/10.48550/arXiv.2409.01754

Zhang, Y., Das, S. S. S., & Zhang, R. (2024). *Verbosity $\neq$ Veracity: Demystify Verbosity Compensation Behavior of Large Language Models* (No. arXiv:2411.07858). arXiv. https://doi.org/10.48550/arXiv.2411.07858

Zhou, L., Schellaert, W., Martínez-Plumed, F., Moros-Daval, Y., Ferri, C., & Hernández-Orallo, J. (2024). Larger and more instructable language models become less reliable. *Nature*, 1–8. https://doi.org/10.1038/s41586-024-07930-y